ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 1,Jan. 2024,pp. 215－222
DOI:10. 16337/j. 1004-9037. 2024. 01. 019

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种面向大规模资源发现的分布式局部聚类方法

孟新宇[1,2]，潘文宇[1]，马艺宁[1]

(1. 江苏警官学院刑事科学技术系,南京 210031;2. 痕迹检验鉴定技术公安部重点实验室(中国刑事警察学院),沈阳 110854)

摘　要：在大规模资源环境下,传统的资源索引机制导致Peer结点数量急剧增加和负载均衡性能下降,影响查询效率和系统稳定性。本文提出了一种质心模型的局部资源聚类方法,通过将相近资源聚类于单一结点并选出代表性键,有效减少了P2P(Peer-to-peer)网络中的Peer结点规模。此外,局部聚类机制集中处理距离相近的键,避免了资源覆盖的过度膨胀。实验结果显示,基于质心模型的Skip Graph算法不仅降低了查询复杂度,提高了负载均衡性能,而且在网络规模、数据量及查询复杂度方面展现出优秀的扩展性,更好地适应大规模资源发现的需求。

关键词：局部聚类;资源发现;P2P网络;质心模型

中图分类号：TP302　　**文献标志码**:A

A Distributed Local Clustering Method for Large-Scale Resource Discovery

MENG Xinyu[1,2], PAN Wenyu[1], MA Yining[1]

(1. Department of Forensic Science, Jiangsu Police Institute, Nanjing 210031, China; 2. Key Laboratory of Impression Evidence Examination and Identification Technology (National Police University of China), Shenyang 110854, China)

Abstract: In large-scale resource environments, traditional resource indexing mechanisms lead to a rapid increase in the number of Peer nodes and a decrease in load balancing performance, affecting query efficiency and system stability. This paper introduces a centroid model-based local resource clustering method, which clusters similar resources at a single node and selects a representative key value, effectively reducing the scale of Peer nodes in the peer-to-peer (P2P) network. Additionally, the local clustering mechanism focuses on processing closely related key values, thus preventing excessive expansion of resource coverage. Experimental results demonstrate that the Skip Graph algorithm based on the centroid model not only reduces query complexity and improves load balancing performance, but also exhibits excellent scalability in terms of network size, data volume, and query complexity, better adapting to the needs of large-scale resource discovery.

Key words: local clustering; resource discovery; peer-to-peer (P2P) network; centroid model

引　　言

随着信息技术广泛应用和飞速发展,存在于互联网上的资源数据规模日益增大。面对大规模的数

基金项目:痕迹检验鉴定技术公安部重点实验室(中国刑事警察学院)资助项目(HJKF201906);江苏警官学院大创项目(2023)。
收稿日期:2023-11-23;**修订日期**:2024-01-09

据查询和资源发现,传统的C/S架构越来越不适用,学术界和工业界开始在分布式技术上寻求解决方法。在此背景下,分布式聚类算法的研究变得尤为重要。这些算法旨在处理分布在多个节点的大量数据集,优化计算和存储资源利用,同时降低网络通信的负担。例如,基于MapReduce的K-means算法,在不同节点上并行处理数据,有效提升处理大规模数据集的能力[1]。此外,共识聚类方法通过合并不同节点上独立执行的聚类结果来达成最终的聚类,这种方法在生物信息学和社交网络分析中表现出色[2-3]。而对于实时数据流,如金融交易数据,流聚类算法如BIRCH和STREAM,能够动态处理持续到达的数据[4]。本文的主要工作集中于分布式环境中的资源发现和索引问题。现有的P2P(peer-to-peer)研究尽管取得了显著的成果,但在处理大规模资源时面临结点数量急剧增加和负载均衡性能降低的问题。基于分布式哈希表(Distributed Hash table, DHT)的方法[5-7]一直被广泛研究,但由于DHT基于哈希算法[8-10]失去了结点的位置和其键之间的关联性,不能有效地支持范围查询。Skip graphs[11]和SkipNet[12]等算法在实现较好的查询效率基础之上,还支持了范围查询。基于Skip graph的相关算法[13-16]各有优劣。基于虚拟服务的负载均衡算法、虚拟服务和目录相结合的系统负载均衡算法,以及根据节点容量确定虚拟节点的数量、使负载与节点的存储容量相适应的算法,并未考虑节点在CPU、内存及带宽等方面的异构性,负载均衡效果并不理想[17-19]。文献[20]是要解决在动态网格环境下多维属性查询的问题,但该系统基于DHT结构构建,范围查询并不能得到支持,同时当数据量变大、查询维数增加时,其查询效率会急剧下降。基于节点能力的入度调整算法[21]容易使较多的负载集中于处理能力强的节点,导致能力强的节点负载过重。文献[22]提出了先由多维空间映射到一维空间,然后基于P2P对一维空间数据存储索引,但对空间是静态划分,很大程度上限制了该算法在属性动态变化的网格资源发现领域的应用。基于Leopard对等网络结构的负载均衡策略利用2个哈希函数进行负载均衡,考虑节点间的临近性,需要对现有的覆盖网协议作大量改动,难于在实际中得到广泛的应用[23]。基于DHT的P2P多维属性查询打破了数据的连续性[24],不能有效地支持范围搜索。注册节点过载时,Sloppy hashing方法[25]会把无法处理的信息“溢出”到其他节点而成为新的注册节点,进而分担过载节点的负载。通过划分树动态地将多维数据空间映射到一维数据空间上的方法存在的主要问题是空间划分数目随着维数的增加成指数增长,因此当维度较大时系统的性能较差[26]。基于K-ary树的P2P负载均衡算法,K-ary负责收集和发布节点的负载信息,生成基于虚拟服务器的转移策略[27],该算法采用转移就近原则,利用界标簇算法计算节点间的物理距离,但比较复杂,存在单点失效问题。Node Wiz方法[28]将属性动态划分成一个树,能够支持范围查询,同时数据的加入以及节点的加入或者离开引起更新需要的消息数量与维数无关,但是树型结构在根节点处会形成数据路由的瓶颈,在动态环境下大量的数据维护信息可能导致系统的瘫痪。

针对上述问题,本文提出了一种新型的质心模型局部资源聚类方法。该方法不仅考虑了节点的异构性,还通过局部聚类机制集中处理距离相近的键,有效避免了资源覆盖的过度膨胀。实验结果表明,基于质心模型的Skip graph算法在降低查询复杂度、提高负载均衡性能方面表现出色,特别是在网络规模、数据量和查询复杂度方面展现出更好的扩展性,能够更有效地适应大规模资源发现的需求。

1 Skip graph 数据结构

Skip graph数据结构采用关系向量机制在保持结点有序的情况下实现了较好的时间复杂度和性能。相比经典的P2P数据结构DHT,其主要优势是支持范围查询和较好的负载平衡。Skip graph本质上是一个分布式的多层链表,是Skip list在分布式环境下的改进版本,其数据查询的主要思想是贪婪算

法。Skip list的结构定义如下:

(1)在最低层(level 0),所有结点组成一个按键有序的双向链表。

(2)当$i>0$时,第$i-1$层的结点以概率P出现在第i层中。

(3)在每一层,结点都保存其左右邻居[29-30]的信息,类似一个双向链表。

(4)定位某个键时,从最高层开始进行顺序查找,如果检索不到则下降到下一层,直到找到目标值或最接近值。

这种结构使得一个Skip list有log$N$层(其中N为结点总数),显然其查询复杂度为$O(\log N)$。Skip list虽然具有较好的查询复杂度,但它并不适合分布式环境。因为在Skip list中level i上的结点以持久的概率P出现在level$(i+1)$中,这样导致上层结点数不断减少,使得查询对高层结点依赖太大,产生热点问题;此外,由于搜索可能开始于任何一个结点,如果情况十分不幸,搜索将在level 0进行,导致$O(N)$的查询复杂度。

Skip graph在此基础上进行了如下改进,从而适应了分布式环境的需求:

(1)在level 0的链表中,所有结点有序分布,组成双向链表。

(2)每个结点具有一个关系向量。关系向量的每一位随机取自字母表Σ(一般地,取$\Sigma=\{0,1\}$)。例如,在图1中结点2的关系向量为"010"。

(3)当$i>0$时,第$i-1$层中关系向量前i位相同的结点出现在第i层的同一个链表。例如,在图1中,结点1、2、3在第1层中组成一个链表,因为它们关系向量的1位前缀都为"0"。

上述改进有效地解决了热点问题,使得Skip graph成为分布式的数据结构。例如,如图1所示,假设结点1要查询结点7,则结点1在其最高层(第2层)开始顺序查找,首先将查询请求交给其右邻居结点3,结点3发现其右侧没有邻居,则下降一层到第1层进行顺序查找,发现其右邻居即为结点7,查找成功。显然,Skip graph的查询时间复杂度是$O(\log N)$。

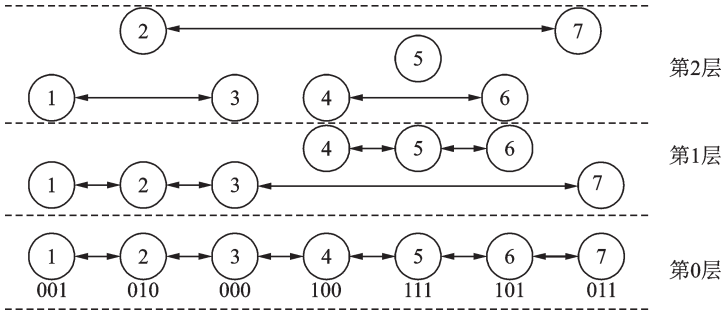

图1 Skip graph算法示意图

Fig.1 Sketch map of Skip graph algorithm

2 质心模型

在Skip graph中,资源索引和Peer结点是一一对应的,这种分布机制在大规模资源环境下显得很不适用,会使得P2P网络中的结点数太多、规模太大,不仅影响查询效率,而且会导致系统的稳定性和容错性降低。所以,文献[31-32]提出一种资源聚类的策略,以防止覆盖资源规模的过度庞大导致结点规模急剧增加。

本文提出了一种代表点的机制,叫作质心模型。将相近资源作为一个分类分布到某个结点上,并

通过合理的方法在这个分类中选择出一个代表点作为该结点的键。代表点机制有利于减少P2P覆盖网中的Peer结点规模,另外通过局部聚类机制使得距离相近的键集中分布,有效地降低了查询时间,提高了检索效率和负载均衡性能。下面给出质心模型的详细算法。

考虑到Skip graph是有序的分布式数据结构,因此质心模型在局部聚类时仍须保持分类以及键的整体有序性。

定义1 $RK_i \prec RK_j$,当且仅当$\forall x \forall y\left(x \in D_i \wedge y \in D_j \rightarrow x \prec y\right)$,其中分类为$D_1, D_2, \cdots, D_n$,每个分类对应的代表点为$RK_1, RK_2, \cdots, RK_n$,定义$\prec$为Skip Graph中key的偏序关系,$x$、$y$为键。

定义2 键唯一数值表示$\mathrm{Uni}(\mathrm{key})=\sum_{i=0}^{\|\mathrm{key}\|-1} d^{S_{\max}-i-1} * \mathrm{Num}(\mathrm{key}[i])$,其中$\|\mathrm{key}\|=\mathrm{length}(\mathrm{key})$,$\mathrm{Num}(\mathrm{key}[i])$表示key第$i$个字符所表示的数字。规定key的字符串最大长度$S_{\max}$,即有$1 \leqslant \|\mathrm{key}\| \leqslant S_{\max}$;$d$表示当前数值为$d$进制数值;$\mathrm{Num}(0) \prec \mathrm{Num}(1) \prec \cdots \prec \mathrm{Num}(a) \prec \cdots \prec \mathrm{Num}(z)$。简单地,令$\mathrm{Num}(0)=0$,$\mathrm{Num}(z)=35$,且不考虑字母大小写,其他字符归0;容易得到$d=36$,键惟一数值是一个36进制的数值。

定义3 相似度函数$\mathrm{Sim}\left(\mathrm{key}_1, \mathrm{key}_2\right)=\dfrac{\left|\mathrm{Uni}\left(\mathrm{key}_1\right)-\mathrm{Uni}\left(\mathrm{key}_2\right)\right|}{\mathrm{Uni}(Z)-\mathrm{Uni}(O)}$,其中$Z$表示长度为$S_{\max}$的字符串"$z\cdots z$",$O$表示长度为$S_{\max}$的字符串"$0\cdots 0$",显然$0 \leqslant \mathrm{Sim}(\mathrm{key}_1, \mathrm{key}_2) \leqslant 1$,并且前缀越相近,两个键越相似。

定义4 所有key组成的集合$D=\{x_1, x_2, \cdots, x_n\}$,并且被划分为$c$个互不重叠的分类$D_1, D_2, \cdots, D_c$;$m_i=\dfrac{1}{\|D_i\|}\sum_{x \in D_i} x$为子集$D_i$的质心。

由上述定义有,任意key与D_i的相似度等价于任意key与D_i的代表点RK_i的相似度$\mathrm{Sim}\left(\mathrm{key}, RK_i\right)$,其中$RK_i$可以不是$D_i$集合中的key值。显然,可以将质心作为代表点。

质心模型在使用质心作为代表点的同时,也为每个分类定义了一个空间范围,即在$[RK-r, RK+r]$内的键将属于同一个分类,其中r表示这个区域的半径。这个机制减少了相似键的查询时间,保证系统的负载均衡。

如上所述,当一个新的键插入到P2P覆盖网中时,需要为这个键找到一个合适的分类插入。这个新插入的键会影响这个目标分类的质心值。RK值将被重新计算,使得这个分类的空间范围也将移动,本文称这种现象为质心漂移,如图2所示。假设系统中存在两个分类D_1和D_2,质心分别是RK_1和RK_2,当一个新的键插入到系统中,并且不能插入到上述两个中的任一个分类中时,则新产生一个分类D_{new},并定义了新分类D_{new}的空间范围,这个新的空间范围包含了原有分类中的键,这样新的分类质心将被重新计算,随之产生空间范围的移动。新的分类空间范围移动,随之也会影响与之重叠的邻居分类的空间范围,例如原有的RK_1和RK_2将重新计算,并且各自分类的空间范围分别向左移动和向右移动。

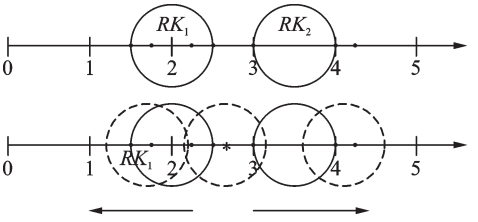

图2 质心漂移

Fig.2 Centroid drifting

每当一个新的键插入、删除和更新时,都将产生质心漂移现象,本文将通过介绍键插入算法来阐述

质心模型如何控制空间范围的移动。由于键的删除过程类似,不再重复介绍。

当一个新的键 key_{new} 插入到系统中,首先获取与此 key_{new} 最相近的代表点 RK_k 和 RK_{k+1},有 $RK_k \prec \text{key}_{\text{new}} \prec RK_{k+1}$。同时计算 key_{new} 与 RK_k、RK_{k+1} 的相似度 $\text{Sim}(\text{key}_{\text{new}}, RK_k)$ 和 $\text{Sim}(\text{key}_{\text{new}}, RK_{k+1})$,下面给出基于质心模型的插入算法。

(1)如果 $\text{Sim}(\text{key}_{\text{new}}, RK_k) \leqslant r_k$ 且 $\text{Sim}(\text{key}_{\text{new}}, RK_{k+1}) > r_{k+1}$,则将这个新的键插入到分类 D_k 中。并且有 $D_k \leftarrow D_k \cup \{\text{key}_{\text{new}}\}$, $m_k \leftarrow \dfrac{((\|D_k\|-1)m_k + \text{key}_{\text{new}})}{\|D_k\|}$, $r_{\text{temp}} \leftarrow \text{Max}\{r_k, \text{Max}\{\text{Sim}(m_k, x_j)\}\}$,其中 $x_j \in D_j$。即又产生以下3种情况:

①当 $m_k > RK_k$,存在两种情况,如果 $m_k + r_{\text{temp}} < RK_{k+1} - r_{k+1}$,意味着更新后的分类 D_k 与 D_{k+1} 不重合,则令 $RK_k \leftarrow m_k$, $r_k \leftarrow r_{\text{temp}}$,如图3所示。否则,为了避免相交,令 $r_k \leftarrow \text{Max}\left\{\dfrac{RK_{k+1} - r_{k+1} - \text{Min}\{x_j\}}{2}, r_k\right\}$, $RK_k \leftarrow RK_{k+1} - r_{k+1} - r_k$。如图4所示。

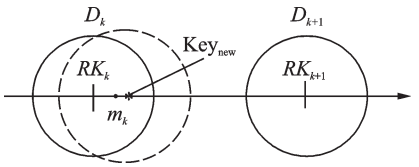

图3 更新后的分类 D_k 与 D_{k+1} 不重合

Fig.3 Updated classification D_k and D_{k+1} not coincide

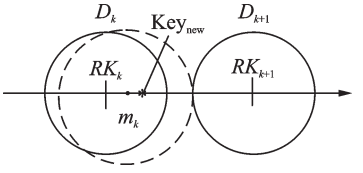

图4 更新后的分类 D_k 与 D_{k+1} 相切

Fig.4 Updated classification D_k and D_{k+1} tangent

②如果 $m_k = RK_k$,则保持 RK_k 和 r_k 不变,如图5所示。

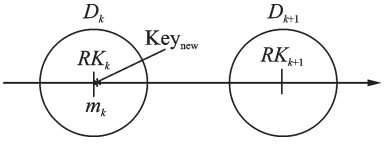

图5 更新后 RK_k 和 r_k 不变

Fig.5 Updated RK_k and r_k unchanging

③如果 $m_k < RK_k$,同样存在两种情况,如果 $m_k - r_{\text{temp}} > RK_{k-1} + r_{k-1}$,则更新后的分类 D_k 将不与 D_{k-1} 相交,则令 $RK_k \leftarrow m_k$, $r_k \leftarrow r_{\text{temp}}$,如图6所示。否则,为了避免相交,则令 $r_k \leftarrow \text{Max}\left\{\dfrac{\text{Max}\{x_j\} - (RK_{k-1} + r_{k-1})}{2}, r_k\right\}$, $RK_k \leftarrow RK_{k-1} + r_{k-1} + r_k$,如图7所示。

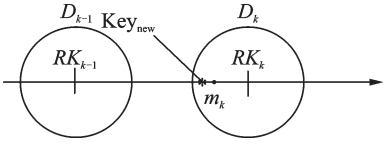

图6 更新后的分类 D_k 将不与 D_{k-1} 相交

Fig.6 Updated classification D_k and D_{k-1} not intersect

(2)对任一 D_k,有 $\text{Sim}(\text{key}_{\text{new}}, RK_k) > r_k$,则新创建一个分类,令 $D_{\text{new}} = \{\text{key}_{\text{new}}\}$, $RK_{\text{new}} = \text{key}_{\text{new}}$, $r_{\text{new}} \leftarrow \text{Min}\{r, RK_{k+1} - r_{k+1} - \text{key}_{\text{new}}, \text{key}_{\text{new}} - RK_k - r_k\}$。其中 r 为系统默认初始半径。图8给出了4种创建不同半径 r 的情况。

根据上述处理过程,分类的半径将迅速膨胀,这意味着更多的键将被插入到这个分类中。为了保持系统负载均衡,本文提供一种对半拆分策略来劈分负载过重的分类。

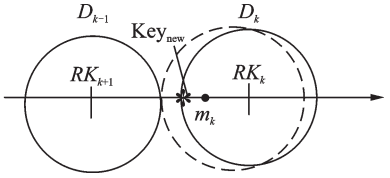

图7 更新后的分类 D_k 与 D_{k-1} 相交

Fig.7 Updated classification D_k and D_{k-1} intersect

(3)如果一个分类的键数量超过阈值 t,则现有的分类被劈分成2个拥有相同键数量的新分类。

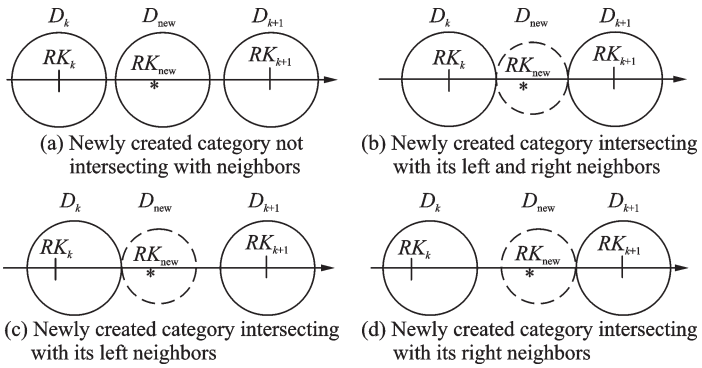

图8 4种创建新分类的情况

Fig.8 Four situations of new creating category

3 实验和分析

一个理想的P2P资源发现系统应该具有高查询效率、低系统成本,同时在网络规模、数据数量和查询复杂度上应提供较好的扩展性。针对上述的性能指标设计了实验环节。实验从Internet上获取了5 000篇文档,由此提取出约17万个键值。本文把这些键值作为索引,分布到改进的Skip graph上。实验中的参数设置如表1所示。

表1 实验参数设置

Table 1 Experimental parameter settings

参数	描述	初始值
m	文档的数量	5 000
n	劈分得到的键值数量	252,630
p	Peer结点数量	252,630
r	质心模型半径大小	1.250 3e−5
T	质心模型中分类的容量阈值	110

由于采用了质心模型,相近的键值被分布到了同一个结点中,从而在一定程度上限制了网络规模的膨胀。因此,对于给定数量的资源,采用质心模型的Skip graph在查询效率上要明显优于普通的Skip graph。图9给出了两种模型在不同文档数下的平均路由跳数的对比。根据质心模型,半径越大,则当前分类的空间范围越大,可能有越多的键放入该分类中。半径越小,极端的情况则倾向于一个分类只拥有一个键,退化为普通的Skip graph。图10给出了在不同半径、不同资源数量下系统中存在的分类数量,即Skip graph中的结点数目。当半径过大,如r=7.563 0e−4时,随着文档不断加入系统,分类数量几乎不变,意味着不断增加的键将加入到原有分类中,而不会增加新的分类。只有当分类中键数量超过系统设定阈值时,才分裂成两个新的子分类,导致分类数量缓慢增加。当半径过小,如r=1.701 1e−7时,随着文档不断加入系统,分类数量快速增加,这是因为由于分类设定的范围空间过小,使得新增加的键很多情况下创建新的分类。显然,半径多大或过小,都会引起系统效率降低,导致系统负载不均衡。通过实验分析得出,当r=1.250 3e−5时,系统具有较好的查询效率和负载均衡。

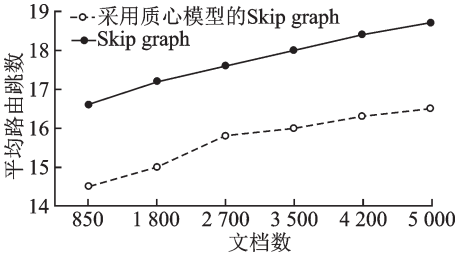

图9 采用质心模型的Skip graph与基本Skip graph平均路由跳数对比

Fig.9 Comparison of average routing hops between the Skip graphs using centroid model and the basic

图11给出了r=1.250 3e−5时不同分类容量在系统中出现的次数,从一定程度上反应了系统的负载均衡。可以看出,大部分的分类容量都处于[60,100]之间,只有极少数分类的容量处于100以上。这

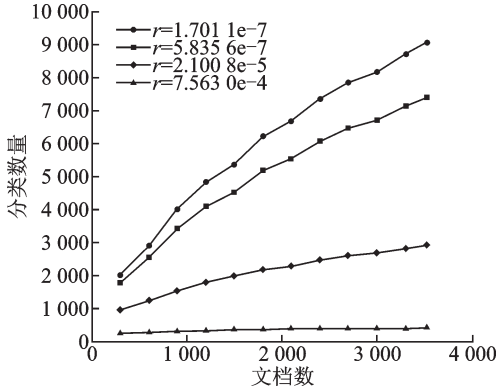

图10 质心模型中不同半径下的平均路由跳数对比

Fig.10 Comparison of average routing hops under different radii in the centroid model

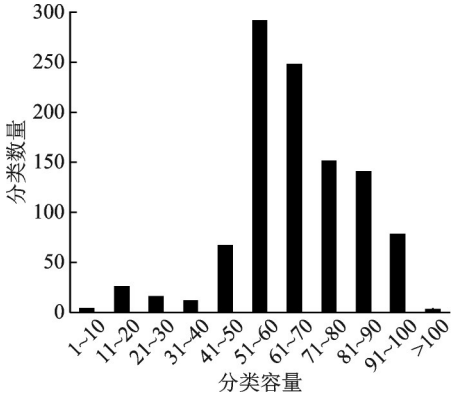

图11 质心模型中分类的负载

Fig.11 Load classified in the centroid model

种局面保证了不会有过热的分类出现,体现了系统的平衡性能。

4 结束语

本文提出了一种针对大规模资源的、基于Skip graph的数据发现模型,重点讨论了改进的Skip graph结构中采用的质心模型,详细阐述了该模型的算法。对原型系统的性能测试表明,采用质心模型的Skip graph获得了较好的查询效率和负载均衡性能。下一步,将在质心模型的基础上研究如何实现资源的分布式多维查询,并且进一步增强系统的容错性能。

参考文献:

[1] MAO Y, GAN D, MWAKAPESA D S. et al. A MapReduce-based K-means clustering algorithm[J]. Supercoput, 2022(78): 5181-5202.

[2] KARASINSKA J M, TOPHAM J T, KALLOGER S E, et al. Altered gene expression along the glycolysis-cholesterol synthesis axis is associated with outcome in pancreatic cancer[J]. Clinical Cancer Research, 2020,26(1): 135-146.

[3] LIU Z, LIU L, WENG S, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer[J]. Nature Communications, 2022,13(1): 816.

[4] XIAO H. BIRCH algorithm and data management in financial enterprises based on dynamic panel GMM test[J]. Cluster Computer, 2019, 22 (S2): 4231-4237.

[5] KAWAGUCHI T, BANNO R, HOJO M, et al. Self-refining Skip graph: Skip graph approaching to an ideal topology[J]. IEEE Annual Consumer Communications Networking Conference, 2017,12(1): 441-448.

[6] NITTI M, GIRAU R, ATZORI L, et al. Trustworthiness management in the social internet of things[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(5): 1253-1266.

[7] FERSI G, LOUATI W, JEMAA M B, et al. Distributed Hash table-based routing and data management in wireless sensor networks: A survey[J]. Wireless Networks, 2013, 19(2): 219-236.

[8] WANG J D, ZHANG T, SONG J K, et al. A survey on learning to hash[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 769-790.

[9] SU Shupeng, ZHONG Zhisheng, ZHANG Chao. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval[C]//Proceedings of International Conference on Computer Vision. Seoul, Korea (South): ICCV, 2019: 3027-3035.

[10] FANG Yixian, ZHANG Huaxiang, REN Yuwei. Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing[J]. Knowlege-Based Systems, 2019, 171(1): 69-80.

[11] ASPNES J, SHAH G. Skip graphs[J]. ACM Transactions on Algorithms, 2007, 3(11): 37-1-37-25.

[12] HARVEY N J A, JONES M B, SAROIU S, et al. SkipNet: A scalable overlay network with practical locality properties[C]//

Proceedings of USENIX Symposium on Internet Technologies and Systems. [S.l.]: USENIX, 2003: 146-157.
[13] PUGH W. Skip lists: A probabilistic alternative to balanced trees[J]. Communications of the ACM, 1990, 33(6): 666-676.
[14] ASPNES J, KIRSCH J, KRISHNAMURTHY A. Load balancing and locality in range-queriable data structures[C]// Proceedings of PODC. [S.l.]: IEEE, 2004: 115-124.
[15] GOODRICH M T, NELSON M J, SUN J Z. The rainbow skip graph: A fault-tolerant constant-degree distributed data structure[C]//Proceedings of SODA. [S.l.]: ACM Press, 2006: 384-393.
[16] ZHANG K, WANG S. LinkNet: A new approach for searching in a large peer-to-peer system[C]//Proceedings of Web Technologies Research and Development—APWeb. [S.l.]: [s.n.], 2005: 241-246.
[17] STOICA I, MORRIS R, LIBEN-NOWELL D, et al. Chord: A scalable peer-to-peer lookup service for internet application [C]//Proceedings of the 2001 Conference on Application (SIFCOMM01). San Diego, USA: ACM Press, 2001: 149-160.
[18] RAO A, LAKSHMINARAYANAN K, SURANA S, et al. Load balancing in structured P2P system[C]//Proceedings of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS03). Heidelberg, Germany: Springer Press, 2003: 68-79.
[19] DABEK F, KAASHOEK M F, KARGER D, et al. Wide-area cooperative storage with CFS[C]//Proceedings of the 18th ACM Principles Symposium on Operating Systems O (SOSP01). Banff, Canada: ACM Press, 2001: 202-215.
[20] CAI M, FRANK M, CHEN J, et al. MAAN: A multi-attribute addressable network for grid information services[J]. Journal of Grid Computing, 2004, 2(1): 3-14.
[21] CASTRO M, COSTA M, ROWSTRON A I T. Debunking some myths about structured and unstructured overlays[C]// Proceedings of the 2nd Symposium on Networked Systems Design and Implementation (NSDI). [S.l.]: ACM Press, 2005: 86-93.
[22] GANESAN P, YANG B, GARCIA-MOLINA H. One torus to rule the mall: Multi-dimensional queries in P2P systems[C]// Proceedings of the International Workshop on the Web and Databases. New York: ACM Press, 2004: 19-24.
[23] CHOU J CY, HUANG T, HUANG K. SCALLOP: A scalable and load-balanced peer-to-peer lookup protocol for high-performance distributed systems[C]//Proceedings of International Symposium on Cluster Computing and the Grid. Taiwan, China: IEEE Press, 2004: 19-26.
[24] RATNASAMY S, FRANCIS P, HANDLEY M, et al. A scalable content, addressable network[C]//Proceedings of the 2001 ACM Annual Conference of the Special Interest Group on Data Communication. San Diego, USA: ACM Press, 2001: 67-76.
[25] FREEDMAN M, MAZIERES D. Sloppy hashing and self-organize clusters[C]//Proceedings of the 2nd International Workshop on Peer-to-Peer Systems. Berkeley, USA: [s.n.], 2003: 45-55.
[26] SHU Y, OOI B C, TAN K L, et al. Supporting multi-dimensional range queries in peer-t o-peer systems[C]//Proceedings of the 5th IEEE International Conference on Peer-to-Peer Computing. New York, USA: IEEE, 2005.
[27] ZHU Y, HU Y. Efficient proximity-aware load balancing for DHT based P2P systems[J]. IEEE Transaction of Parallel and Distributed Systems, 2005, 28(4): 46-55.
[28] BASU S, BANERJEE S, SHARMA P, et al. Node Wiz: Peer-to-peer resource discovery for grids[C]//Proceedings of IEEE/ ACMGP2PC'05. Cardiff, UK: IEEE, 2005.
[29] YANG Y, ZHA Z J, GAO Y, et al. Exploiting web images for semantic video indexing via robust sample-specific loss[J]. IEEE Transactions on Multimedia, 2014, 16(6): 1677-1689.
[30] BOIMAN O, SHECHTMAN E, IRANI M. In defense of nearest-neighbor based image classification[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2008: 1-8.
[31] SHENG Weiguo, LIU Xiaohui. A genetick-medoids clustering algorithm[J]. Journal of Heuristics, 2006, 12(6): 447-466.
[32] DHARMENDRA S M, SCOTT S. Feature weighting in K-means clustering[J]. Machine Learning, 2003, 52(3): 217-237.

作者简介:

孟新宇(1977-),**通信作者**,男,博士,讲师,研究方向:智能控制、计算机仿真、刑事技术,E-mail: mengxinyu@jspi.cn。

潘文宇(2004-),男,本科生,研究方向:智能计算。

马艺宁(1991-),男,博士,副教授,研究方向:能量存储、信息技术、刑事技术,E-mail:mayining@jspi.cn。

(编辑:张黄群)